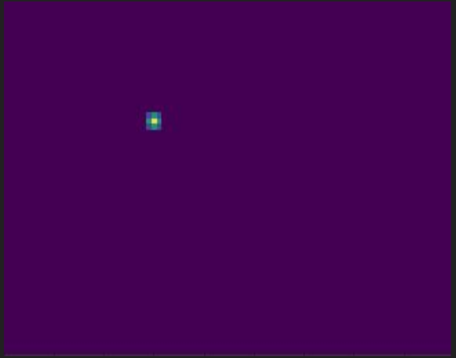
Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation

David Abraham
Simon Dufort-Labbé

Introduction

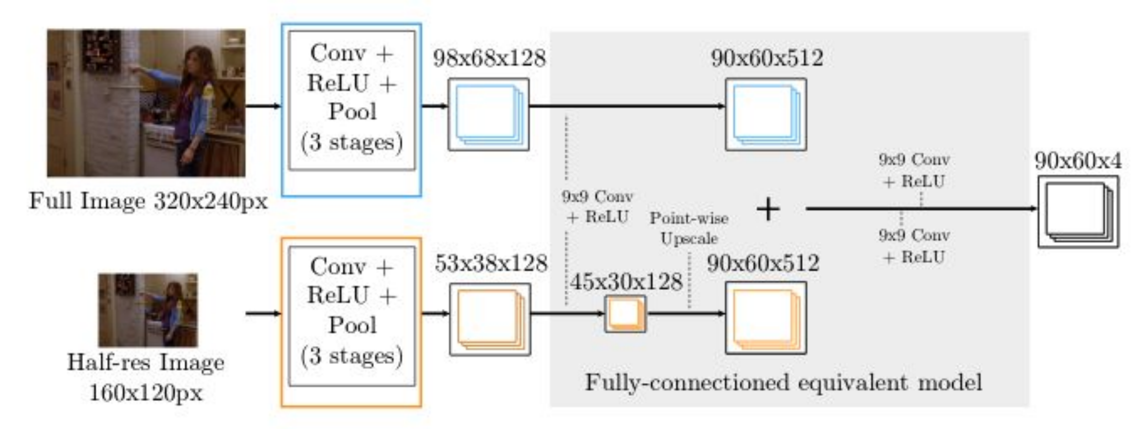
- Implementation based on paper : ”Joint Training of a Convolutional Network and a Graphical Model for Human Pose Estimation” (Tompson et al., 2014).
- Goal : body part locations from a single 2D image
- Model is in two parts :
  - CNN body part detector
  - Graphical spatial model
- Both models are jointly trained at the same time

Dataset



- Frames Labeled in Cinema (FLIC)
- 5003 images (20% for testing)
- 720 x 480 pixels RGB images
- 90 x 60 pixels heatmap for each joints
- Only a single person labeled per image
- Joints labeled : shoulders, elbows, wrists, hips, nose and torso

Model - CNN part detector



- Fully convolutional network
- Multiple inputs : same image at different resolutions
- Output : heatmap for each of the predicted joints

Model - Graphical spatial model

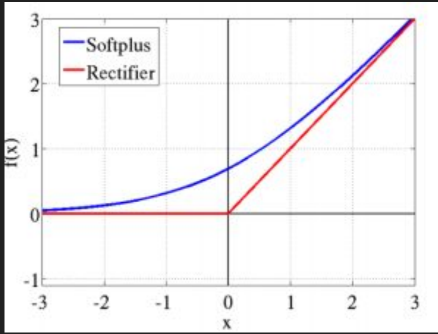
MRF like model, representing a fully connected graph (including self-connection)

$$P_{\bar{A}} = \frac{1}{Z_2} \prod_{v \in \mathcal{V}} p_{A|v} * p_v + b_{v \to A}$$

Proposed estimation

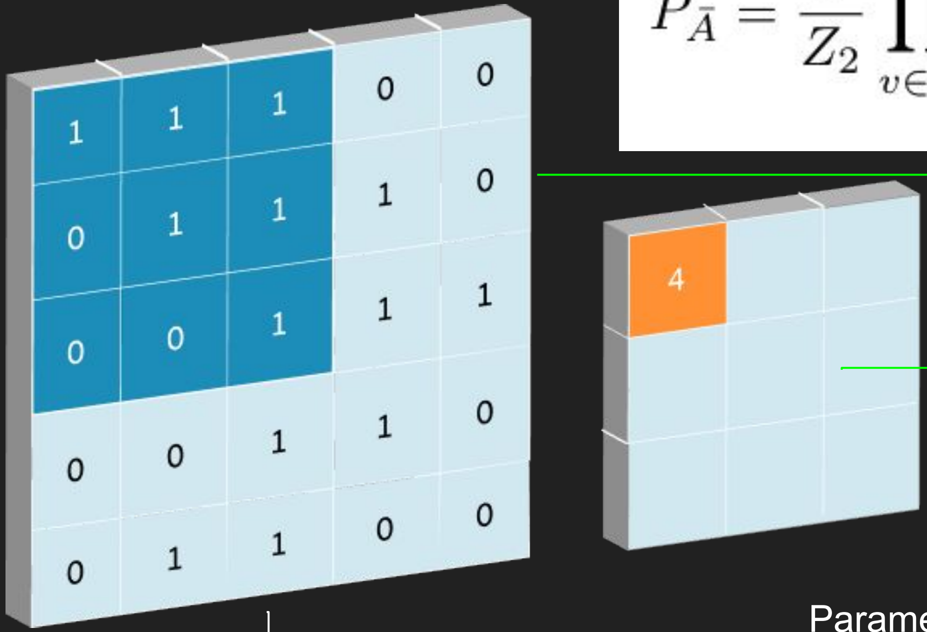
$$\bar{e}_A = \exp\left(\sum_{v \in \mathcal{V}} \left[\log\left(\mathrm{SoftPlus}(e_{A|v}) * \mathrm{ReLU}(e_v) \right) + \mathrm{SoftPlus}(b_{v \to A}) \right) \right] \right)$$

$$\text{where}: \quad \mathrm{SoftPlus}(x) = \frac{1}{\beta} \log(q + \exp(\beta x))$$



Quick Note : convolution over likelihood

$$P_{\bar{A}} = \frac{1}{Z_2} \prod_{v \in \mathcal{V}} p_{A|v} * p_v + b_{v \to A}$$



The resulting product of the likelihood and the prior is learned as the result of a convolution from an inference map and the output from the CNN part detector.

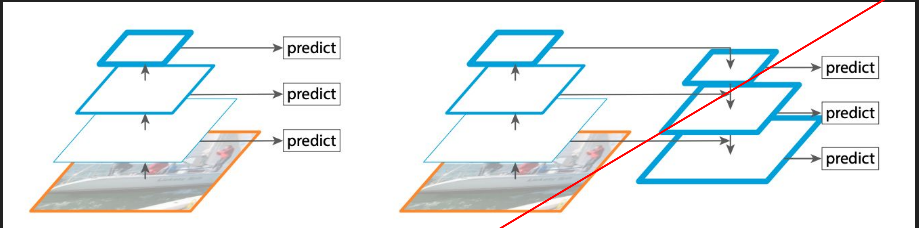
Parameters of this map are learned : they allow the model to learn a probability distribution of a joint given another joint

Changes and improvements

Global change : Train end-to-end from the beginning, cross-entropy instead of MSE

CNN — Part detector:

- Used batch normalization to stabilize training
- Incorporated 3 resolutions scaling instead of two, using a Features pyramidal network inspired approach
- Upscaling and downscaling was done by learned convolution layers.



(c) Pyramidal feature hierarchy (d) Feature Pyramid Network

Spatial model :

$$\bar{p}_A \approx \mathrm{Softmax}\left(\sum_{v \in \mathcal{V}} \left[\log\left(\mathrm{Conv2d}\left[\mathrm{SoftPlus}(p_{A|v}) \ * \ \mathrm{SoftPlus}(p_v)\right] + \mathrm{SoftPlus}(b_{v \to A})\right)\right]\right)$$

No “pre-initialization”

Biggest difference : Use of normalization

- Advantages:
  - More precise predictions
  - Ability to fill a “hole”
- Drawback:
  - Learned the dataset bias : Making predictions for only one group of joints (one person)

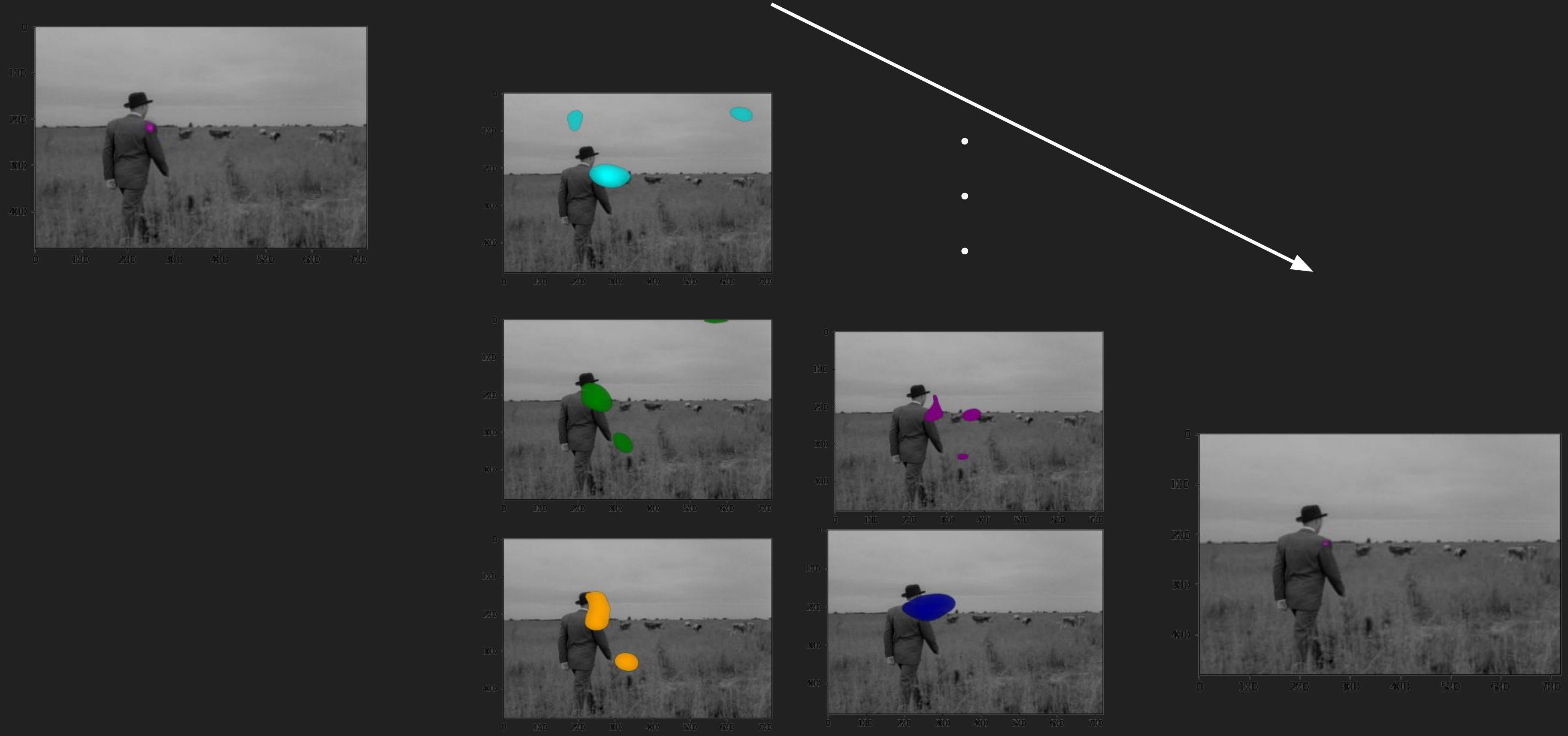
Joints location predictions

CNN only

Full model

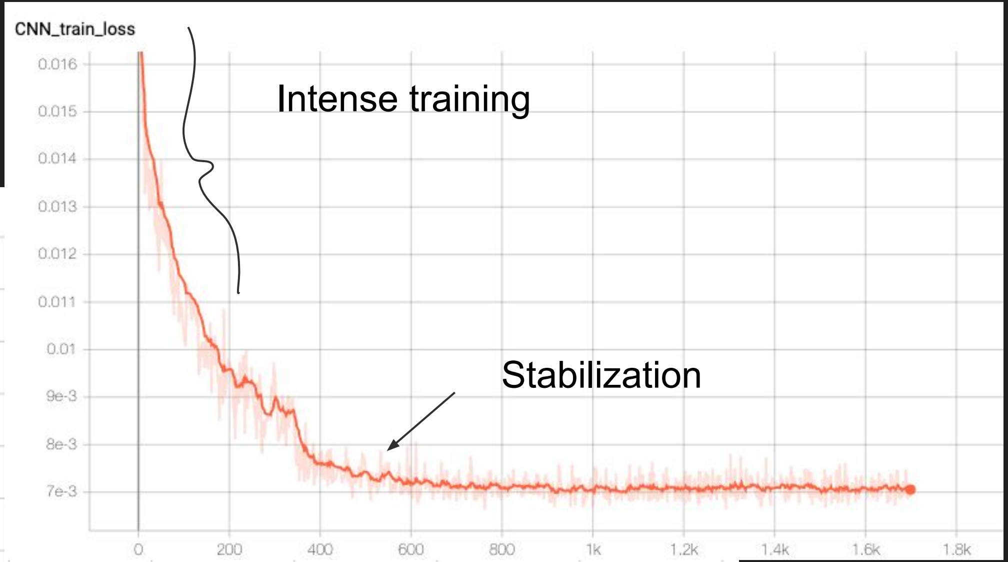
Targets

Visualizing the likelihoods from the spatial model

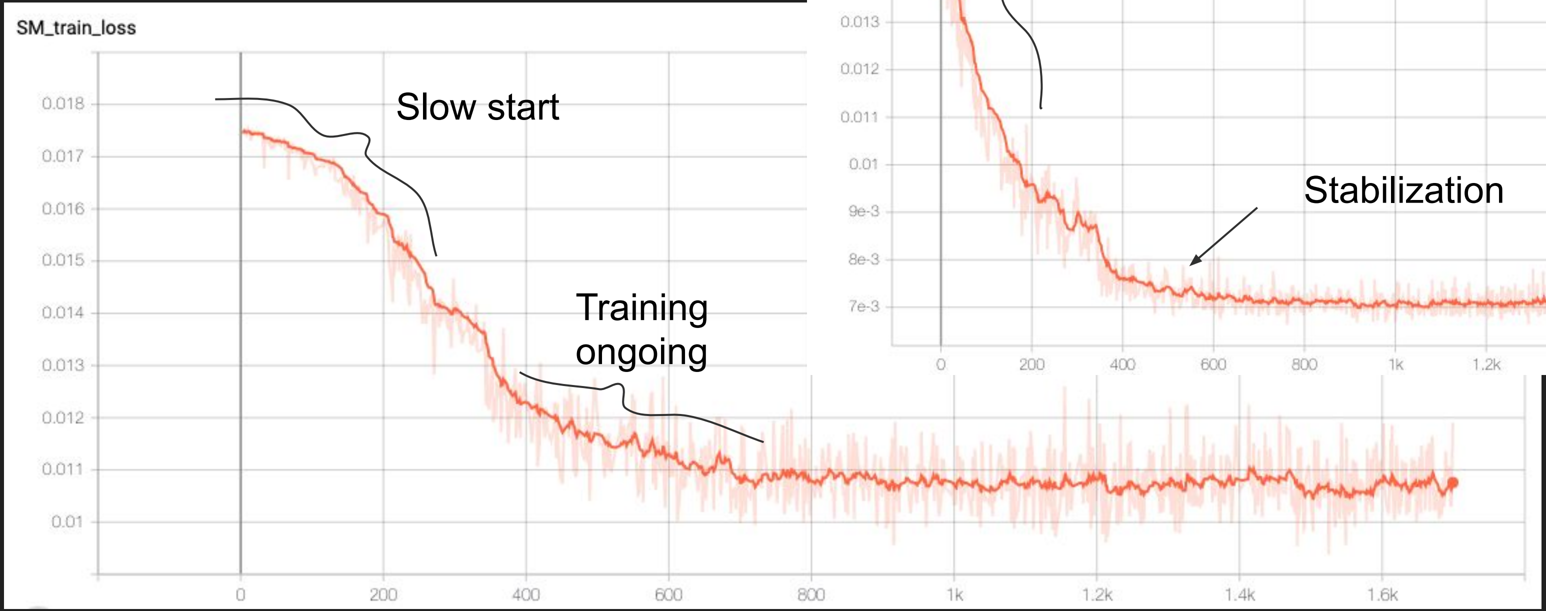


Training process

Loss over the CNN part only



Loss after spatial model correction



Conclusion

- The introduction of a spatial model that, reinforcing the model ability to make structured prediction, can definitely improve the predictions quality of joints
- The proposed spatial model is pretty good at learning the structure between predictions, but will of course incorporate any bias present in the training dataset
- The addition of such a model do not require much more elbow grease, as it is trainable from scratch in an end-to-end fashion

Consideration:

Due to time and computational constraints, we couldn’t check if the addition of a spatial model to a CNN with more capacity (example : ResNet) brought as much amelioration.

Resnet-like models are able to leverage bigger context information, thus already incorporating a form of structured prediction.